Opinion TRENDS in Cognitive Sciences Vol.not known No.not known Month 0000 1

3-D structure perceived from dynamic information: a new theory

Fulvio Domini[1] and Corrado Caudek[2]

[1]Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912-1978, USA
[2]Department of Psychology, University of Trieste, Trieste, Italy

Image movement provides one of the most potent two-dimensional cues for depth. From motion cues alone, the brain is capable of deriving a three-dimensional representation of distant objects. For many decades, theoretical and empirical investigations into this ability have interpreted these percepts as faithful copies of the projected 3-D structures. Here we review empirical findings showing that perceived 3-D shape from motion is not veridical and cannot be accounted for by the current models. We present a probabilistic model based on a local analysis of optic flow. Although such a model does not guarantee a correct reconstruction of 3-D shape, it is shown to be consistent with human performance.

To perceive the 3-D shape of objects from two-dimensional (2-D) retinal images, our brain uses binocular as well as monocular cues such as motion, occlusion, texture and shading [1]. Although 3-D shape perception normally depends on the simultaneous presence of many sources of depth information, human observers show a remarkable ability to extract the 3-D structure of an object from motion cues alone. This ability is referred to as structure-from-motion (*SfM*) perception. The importance of motion for 3-D shape perception was demonstrated 50 years ago by Wallach and O'Connell [2]: in their classic demonstration, the shadow of a wire-frame figure appears flat when the wire frame is stationary, but pops out in depth as soon as the wire frame is rotated.

Currently, the generally accepted theory of *SfM* perception is formulated in terms of the inverse-optics problem: given a 2-D image, the observer needs to determine the 3-D object from which the image is a projection [3–5]. The inverse-optic problem does not have a unique solution: the same 2-D image, in fact, is consistent with infinite 3-D objects. Nevertheless, computer-vision algorithms have been devised that recover the 'correct' 3-D shape (if opportune assumptions are met), either up to scale factor [6–9], or up to a stretching in depth [10], and these algorithms have been proposed as models of human performance [6,11].

Although it is widely believed that human *SfM* perception is veridical (i.e. that observers' perceptions correspond to faithful copies of the projected objects), we have recently shown that perceived 3-D shape from motion is distorted by more than a scale factor or a stretching in depth [12,13] and, therefore, cannot be accounted for by the computational algorithms that have been developed for machine vision. We propose a probabilistic model that, in general, does not guarantee a correct solution to the inverse-optic problem but, nevertheless, is consistent with both veridical performance and perceptual biases in many *SfM* tasks [14,15]. Before presenting its rationale, we will briefly describe the two classes of *SfM* models that have been proposed so far.

Euclidean and affine descriptions

The *SfM* models can be divided into two classes: Euclidean and affine models. Euclidean *SfM* algorithms recover an exact copy (up to a scale factor) of the projected objects [3–5]. If the projected object is a wedge, for example, then the recovered shape will also be a wedge with the same angle between its two planar surfaces as the wedge that generated the image. Euclidean *SfM* algorithms correctly recover the metric properties of the projected 3-D structures. For the simple example of Figure 1, these properties correspond to the slants of the two surfaces (i.e. the tangents of the angles α_0 and α_1 between the surfaces of the wedge and the frontal-parallel plane) and to the absolute distance between any pair of points (e.g. the distance between the points P_0 and P_1 in Figure 1).

From the same image projections, an affine *SfM* algorithm would only recover a copy of the projected wedge that is stretched in depth by an arbitrary amount [10]. In this case, the angle between the two planar surfaces, in general, would not be preserved. Affine geometry provides a more abstract representation of 3-D shape than Euclidean geometry, and affine *SfM* algorithms correctly recover only properties such as the depth order between pairs of points, the parallelism between lines defined by pairs of points on the surface, and the coplanarity among points.

Optic flow

To understand human *SfM* we need also to know what information observers use to derive 3-D shape from motion. An answer to this question will be attempted in the present and the next section.

The image transformations produced by the relative motion between an observer and an object like that of Figure 1 can be described in terms of the 2-D motions of discrete texture elements belonging to the 3-D object. The

Corresponding author: Fulvio Domini (Fulvio_Domini@Brown.edu).

 1364-6613/$ - see front matter doi:10.1016/j.tics.2003.08.007

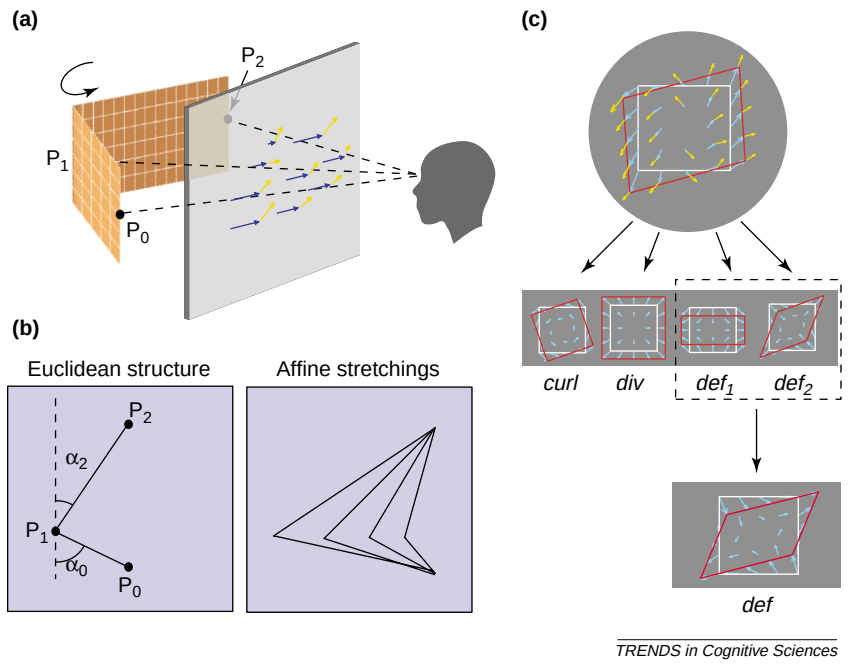

Figure 1. (**a**) A 3-D rotating wedge produces an optic flow on the image plane. Such optic flow is represented by the blue and yellow vectors that correspond to the 2-D velocities of the projected features in two successive instants of time. In a coordinate system in which the *z*-axis corresponds to the line of sight, the *z*-depth distance between P_0 and P_1 is equal to the depth distance between P_1 and P_2. (**b**) View from above the wedge (left panel). Δz represents the depth distance between P_0 and P_1 and between P_1 and P_2; x_{01} is the size of the projection on the image plane of one of the two surfaces of the wedge; α_0 and α_1 are the angles between each planar surface of the wedge and the frontal-parallel plane. Euclidean *SfM* models reconstruct a 3-D shape from the optic flow in which the angles α_0 and α_1 are derived in a veridical manner. In an affine algorithm (right panel) the wedge is stretched along the *z*-axis by four different amounts. In terms of an affine description, all these structures are equivalent. Consequently, in general, affine *SfM* algorithms do not derive the angles α_0 and α_1 in a veridical manner. (**c**) Motion of the projected texture elements can be described by the instantaneous velocities (blue arrows) and their change over time. This change over time can be measured by considering the values of each velocity vector in two successive instants of time (blue and yellow arrows). If we consider a small patch of the projected surface, its shape undergoes a non-rigid image transformation in two successive moments of time. For example, if this patch projects as a square in the first moment of time, successively the projected patch can take on the shape indicated in red. This 2-D shape change can be decomposed in four elementary transformations: a rigid image rotation (*curl*), an isotropic expansion or contraction (*div*), and two components of shear along vertical (def_1) and oblique axes (def_2). The two shear components quantify the amount of shape change and can be summarized by a unique value called *def*.

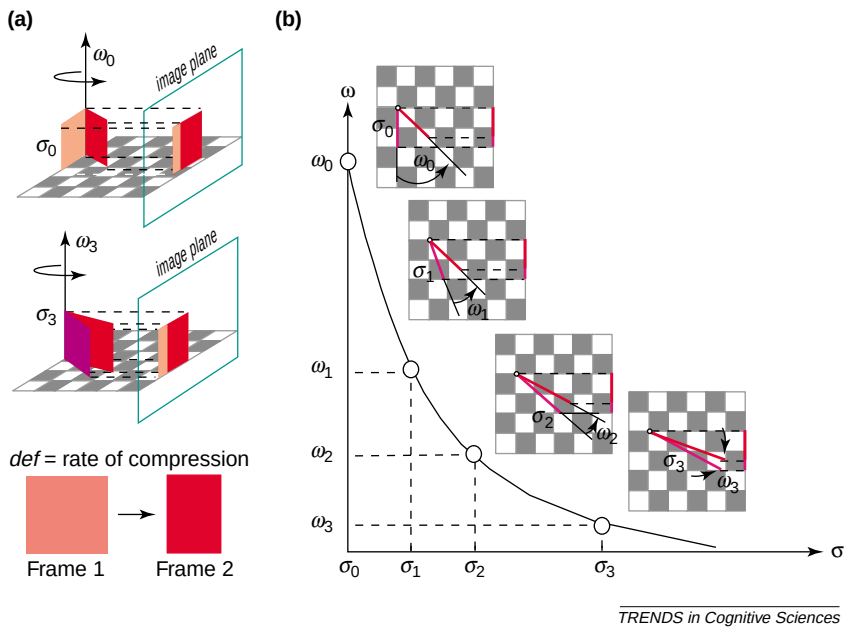

Figure 2. (**a**) Image transformations produced by the vertical rotation of a planar surface. Initially, the 3-D surface projects a square on the image plane (pink). After rotation of the planar surface, the projected square is compressed (red). The rate of this compression represents the amount of *def*. Note that a large rotation of a surface having a small initial slant (top) could produce exactly the same amount of compression as a small rotation of a surface with an initial larger slant (bottom). (**b**) Family of slant and angular rotation magnitudes that produce the same image compression (*def*). Four instances of these σ, ω pairs are represented by the view from above the rotating surface.

overall motion of these feature elements is called optic flow [16]. For the present purposes, it is important to distinguish between a description of optic flow in terms of instantaneous velocities (the so-called first-order properties of the flow), and one in terms of velocities and accelerations (second-order properties) [17].

If the object in Figure 1 moves rigidly relative to a stationary observer, then a veridical description of its full Euclidean structure can, in principle, be derived from second-order optic flow [6,7]. The first-order properties of optic flow are sufficient, conversely, to specify only the affine structure of this object (i.e. its 3-D shape up to a depth stretching) [10].

Many psychophysical findings have demonstrated that the judgments of Euclidean properties from optic flow are not veridical [11,18,19]. In particular, it has been shown, first, that judgments of Euclidean properties are far less accurate than judgments of affine properties [11], and second, that human observers have a very limited sensitivity for second-order temporal properties and thus rely mainly on velocities to recover 3-D information from optic flow [11,20].

Initially these two findings prompted researchers to hypothesize that the perceptual analysis of optic flow is restricted to a veridical recovery of the affine properties of the projected objects [11]. In a series of investigations, however, we have also shown that the affine properties are not recovered from optic flow in a veridical fashion [12,13,21]. These findings have thus led to the conclusion that the perceptual analysis of optic flow cannot be accounted for by either Euclidean or affine *SfM* algorithms [13]. To account for these findings, we have developed a model that derives the local orientation and motion of 3-D surface patches from an ambiguous property of first-order optic flow called deformation [14,15]. In the next section we will describe this property and its relation to the parameters of the projected 3-D shape.

The ambiguity of local optic flow

First-order optic flow can be described intuitively by representing the retinal projections with a sequence of frames, and by considering how a local patch of the projected surface is distorted from one frame to the next. Such distortions can be thought of as a sum of four elementary changes (see Figure 1c). Two of them modify only the orientation (‘*curl*’) and the size (‘*div*’) of the projected patch; the other two modify the shape of the 2-D projection and are therefore called shearing or deformation. The two shearing components are the only image transformations that are informative about 3-D shape and they can be summarized by a single quantity called *def* [17].

For the sake of simplicity, let us consider the optic flow produced by the rotation about the vertical axis (ω) of one of the two planar surfaces of the wedge represented in Figure 1. After the rotation, the initial projection of this surface on the image plane will be compressed. In this particular case, *def* coincides with the amount of compression (Figure 2a).

How is *def* related to the metric properties of the projected 3-D structures? It should be noted that the

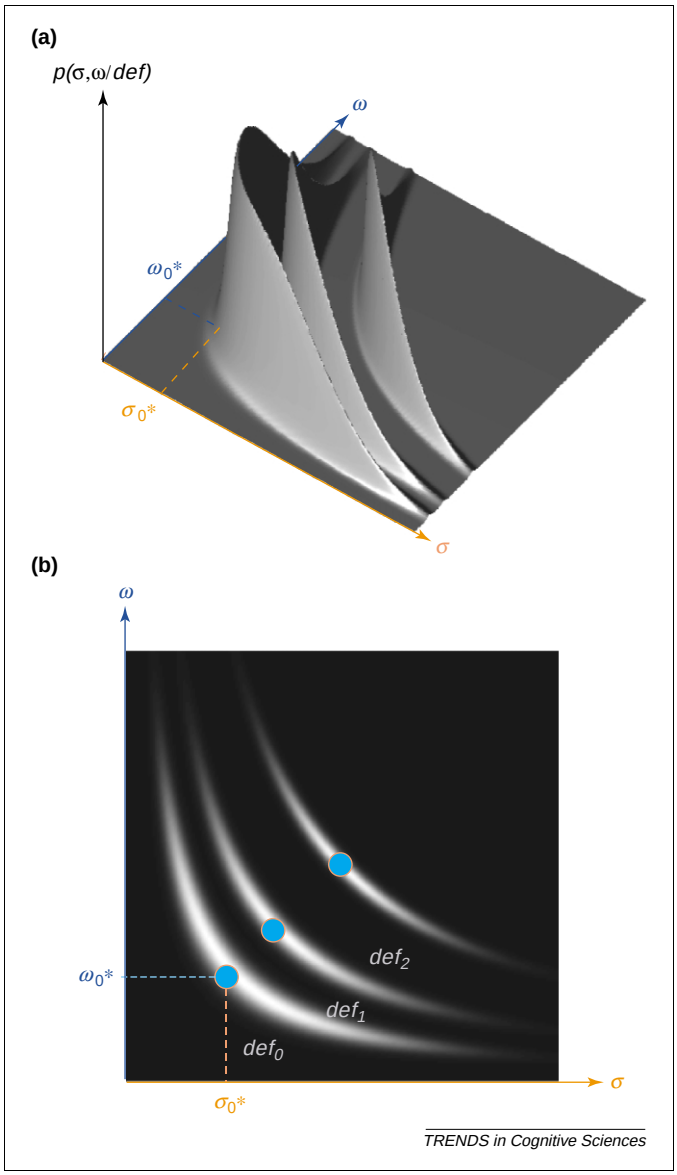

Fig. 3. (**a**) Probability density functions $p(\sigma, \omega \mid def)$ for three different *def* values. These probability functions have been calculated by assuming Gaussian noise in the measurement of *def*. The point $\sigma * \omega *$ represents the most likely σ, ω pair given *def*. (**b**) The probability functions of (a) are coded by gray levels in which white represents the highest value and corresponds to the most likely σ, ω pair given *def* (red-blue dots).

compression produced by, say, a small slant and a large angle of rotation produces the same deformation as a small rotation of a more slanted surface (see Figure 2). There are infinite combinations of slant and angular rotation, therefore, which give rise to the same *def* (Figure 2) [14]. In the instantaneous and local case (i.e. for two frames very close in time and for a very small patch), *def* represents the rate of compression (or shearing, in the general case), and it is related to the slant σ and the angular velocity ω by:

$$def = \sigma\omega \quad (1)$$

Functional MRI studies have identified in the middle temporal area (MT/V5) and in the fundus of the superior temporal sulcus (FST), the human brain areas that have a basic role in *SfM* processing [22–24]. Moreover, neural activity in the dorsal division of the medial superior temporal

Box 1. Distortions of perceived depth and shape

Another important characteristic of our model is that it derives perceived slant as a non-linear function of *def* (see Eqn 2 in main text). What does this imply for the judgments of relative depth? Consider the probe dots P_0 and P_1 in Figure Ia. From simple trigonometry, the depth difference $\Delta z = P_0 - P_1$ is equal to $\Delta z = \Delta x\sigma$, where Δx is the size of the planar patch projected onto the image plane and σ (slant) is the tangent of the angle α. Accordingly, the perceived depth separation $\Delta z'$ is equal to $\Delta z' = \Delta x\sigma'$. The relation between the perceived and the simulated depth difference, therefore, can be expressed in terms of the ratio between the two previous equations:

$$\frac{\Delta z'}{\Delta z} = \frac{\sigma'}{\sigma} \quad (3)$$

If perceived slant is expressed in terms of Eqn 2, we obtain

$$\sigma' = \sqrt{\frac{\sigma_{max}}{\omega_{max}}}\sqrt{\sigma\omega} = k\sqrt{\sigma} \quad (4)$$

From Eqn 3 and Eqn 4, moreover:

$$\Delta z' = \frac{k}{\sqrt{\sigma}}\Delta z \quad (5)$$

Eqn 5 reveals that, according to our model, the perceived depth difference $\Delta z'$ between P_0 and P_1 should be proportional to the simulated depth separation between the two dots (Δz), and inversely related to the slant (σ) of the patch on which P_0 and P_1 are located. This prediction was confirmed by Domini and Braunstein [13].

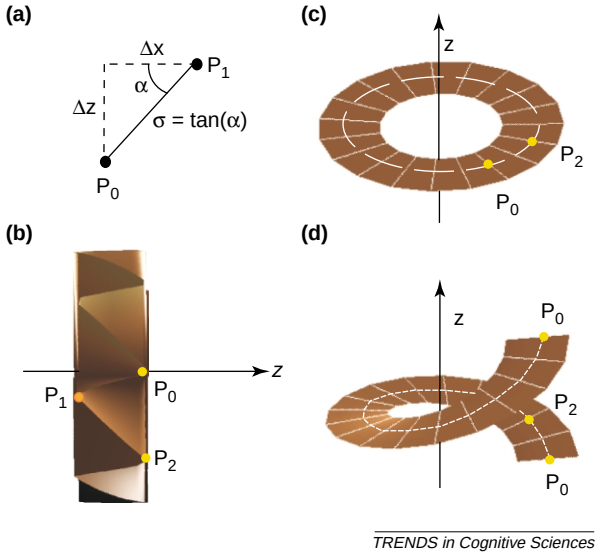

Figure I. (**a**) Top view of a planar surface slanted away from the image plane by an amount $\sigma = tan(\alpha)$. Δx is the horizontal component of the image-plane projection of the surface. Δz is the *z*-axis distance between the points P_0 and P_1. (**b**) Side view of the circular crown surface used by Domini *et al.* [12] to investigate internal consistency of depth judgments in structure-from-motion. The *z*-axis represents the line of sight, and Δz is the *z*-depth distance $P_0 - P_1 = P_2 - P_1$. (**c**) Schematic representation of the plane defined by the points P_0 and P_2 of each wedge element of the circular crown. In successive trials, the probe dots P_0 and P_2 were located in different positions along the circular path indicated by dashed white line. Even if the probe dots P_0 and P_2 were always simulated at the same distance from the observer, because P_0 was perceived to be closer than P_2 for each wedge element composing the circular crown, the integration of the observers' judgments produced the internal inconsistency schematically represented in **(d)**.

Implications for perceived 3-D shape

Our model predicts that local judgments of perceived 3-D shape also violate another important property of real 3-D structures: internal consistency. Let us consider a 3-D structure made up of a circular arrangement of wedges, as shown in Figure Ib (this circular crown is composed of elements similar in shape to the wedge represented in Figure 1 in the main text). Suppose one moves along a closed path on the image-plane projection of this surface: each 2-D step along this path corresponds to a depth difference on the 3-D surface. In order for this 3-D shape to be internally consistent, the sum of the depth differences associated with the successive steps along this closed path, from point P_0 back to itself, must be equal to zero. According to our model, however, the perceived depth difference between the dots P_0 and P_1 should be smaller than the perceived depth difference $P_2 - P_1$, because the slant (σ_0) of the patch on which (P_0, P_1) are located is larger than the slant (σ_1) of the patch on which (P_2, P_1) are located. According to our model, therefore, P_2 should be perceived to be closer to the observer than P_0.

To test this prediction, we asked observers to make judgments of depth order along a closed path for the probe dots P_0 and P_2 located at the joints of the planar surfaces making up a circular-crown surface [12], similar to that shown in Figure Ib. As predicted by Eqn 5, the point P_2 was judged to be closer than P_0. The depth judgments along this closed path, therefore, can be thought of as an 'always ascending' staircase, an obvious violation of the property of internal consistency which would result in the 3-D shape represented in Figure Id. This result, together with other similar findings [13], indicate that neither Euclidean nor affine geometries provide an adequate description of perceived 3-D shape from motion.

area (*MSTd*) in the macaque has been found to be specifically sensitive to the *def* component of optic flow [25,26].

Probabilistic interpretation of a moving retinal image

Having identified in first-order optic flow the information that observers use to derive 3-D shape from motion, we need now to describe how perceived 3-D shape can be derived from *def*. Here we will summarize the model that we have proposed, which tries to account for both correct judgments and biases in human *SfM* [14]. This model is based on a local analysis of first-order optic flow and its aim is to find the most likely σ, ω pair given *def* (Figure 3). Our model analyzes the visual motion that is generated by an object moving relative to a passive observer and, thus, it does not take into consideration the specific contributions of extra-retinal information to 3-D shape perception [27–29]. The rationale of the model can be exemplified in the following manner.

Let us suppose that σ and ω are normalized, so that they can vary in the range 0–1 (Figure 3). If *def* is measured without noise, then all the possible σ, ω pairs whose product is equal to *def* lie on a hyperbola (Figure 2b). If the measurement of *def* is perturbed by Gaussian noise, then the 2-D probability density function of σ, ω given *def* is not uniform and has a center of mass corresponding to the point of the hyperbola that is closest to the origin of the coordinate system. The center of mass of this probability density distribution, therefore, corresponds to the point $\sigma = \omega = \sqrt{def}$. In general, if σ and ω vary in the ranges $[0, \sigma_{max}]$ and $[0, \omega_{max}]$ respectively, then the posterior probability distribution of a σ, ω pair given

def has a maximum value corresponding to:

$$\sigma' = \sqrt{\frac{\sigma_{max}}{\omega_{max}}}\sqrt{def} \quad \omega' = \sqrt{\frac{\omega_{max}}{\sigma_{max}}}\sqrt{def} \qquad (2)$$

According to our proposal, then, observers interpret first-order local optic flow by choosing the most likely 3-D solution – that is, the slant and angular velocity magnitudes that maximize the posterior probability of a σ, ω pair given *def*.

To test our model, we simulated the optic flow produced by the rotation of a planar random-dot surface and asked observers to judge perceived surface slant and angular rotation magnitudes [14]. Consistent with the predictions of the model, we found that perceived slant and angular rotation were an increasing function of *def*, whereas the influence of simulated σ and ω was negligible.

The rigidity assumption is not biologically plausible

One of the fundamental assumptions of both Euclidean and affine models is that 3-D objects undergo a rigid rotation unless otherwise specified by the projected image transformations [6,30,31]. We proved that this assumption (one of the fundamental constraints embedded in the algorithms deriving 3-D shape from motion) is not biologically plausible [21].

Let us consider again the wedge in Figure 1. The two planar surfaces of this wedge have different orientations (i.e. slants). If this 3-D structure is rotated rigidly with respect to the observer, then these two planar surfaces will project different deformations (Eqn 1, above). If, according to Eqn 2, the amount of perceived rotation for each surface is derived from *def*, then, in general, a reliable discrimination between rigid and non-rigid motion would not be possible.

This prediction was confirmed by one of our investigations. We asked observers to discriminate rigid from non-rigid motion in stimulus conditions in which two surfaces were rotated rigidly or non-rigidly, and projected either the same or different deformations [21]. The results showed that perceived rigidity was determined by the distribution of *def* and not by simulated rigidity. By means of Eqn 2, moreover, we were also able to account for misperceptions in the orientation of axis of rotation [32], the discrimination between constant and non-constant 3-D angular velocity [33], and the segmentation of two overlapping velocity fields [34]. A further characteristic of our model, which results on distortions of perceived depth, and implications for perceived shape are discussed in Box 1.

Concluding remarks

Even in the absence of other depth information, optic flow is an effective depth cue from which 3-D information of the observed scene can, in principle, be derived. This article has reviewed two hypotheses regarding observers' interpretation of optic flow. One hypothesis is that the perceptual analysis of optic flow is global and veridical according to the 'inverse-optic' approach, as suggested by most of the models of perceived *SfM* that have been proposed thus far. A second hypothesis is that observers analyze optic flow in a heuristic and patch-way fashion. According to this second hypothesis, the most likely 3-D interpretation is assigned to local (ambiguous) first-order properties of the flow and, consequently, the perceptual interpretation is, in general, not veridical, nor internally consistent. The empirical evidence indicates that both veridical judgments and biases in the local judgments of optic flow can be accounted for by this heuristic hypothesis. Rather than solving the 'inverse-optic' problem, therefore, the visual system seems to rely on probabilistic interpretations, perhaps derived from learning. Because the perceptual interpretation of optic flow is not veridical, the goal for future research is to understand how local information is integrated through space and time to achieve a global and coherent perceived 3-D shape [35–37].

References

1 Proffitt, D.R. and Caudek, C. (2002) Depth perception and perception of events. In *Experimental Psychology: Handbook of Psychology* (Healy, A.F. and Proctor, R.W., eds), pp. 213–236, Wiley
2 Wallach, H. and O'Connell, D.N. (1953) The kinetic depth effect. *J. Exp. Psychol.* 45, 205–217
3 Todd, J.T. (2002) The visual perception of 3D structure. In *The Handbook of Brain Theory and Neural Networks* (Arbib, M.A., ed.), MIT Press
4 Todd, J.T. (1998) Theoretical and biological limitations on the visual perception of three-dimensional structure from motion. In *High-Level Motion Processing: Computational, Neurophysiological and Psychophysical Perspectives* (Watanabe, T., ed.), pp. 359–380, MIT Press
5 Todd, J.T. *et al.* (1999) The intrinsic geometry of perceptual space: its metrical, affine and projective properties. In *Fechner Day 99: The End of 20th Century Psychophysics* (Killeen, P.R. and Uttal, W.R., eds), pp. 169–175, The International Society for Psychophysics, Tempe AZ
6 Ullman, S. (1979) *The Interpretation of Visual Motion*, MIT Press
7 Longuet-Higgins, H.C. and Prazdny, K. (1980) The interpretation of a moving retinal image. *Proc. R. Soc. Lond. B. Biol. Sci.* 208, 385–397
8 Bennett, B.M. *et al.* (1989) Structure from two orthographic views of rigid motion. *J. Opt. Soc. Am. A* 6, 1052–1069
9 Ullman, S. (1984) Maximizing rigidity: the incremental recovery of 3-D structure from rigid and nonrigid motion. *Perception* 13, 255–274
10 Koenderink, J.J. and van Doorn, A.J. (1991) Affine structure from motion. *J. Opt. Soc. Am. A* 8, 377–385
11 Todd, J.T. and Bressan, P. (1990) The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Percept. Psychophys.* 48, 419–430
12 Domini, F. *et al.* (1998) Distortions of depth-order relations and parallelism in structure from motion. *Percept. Psychophys.* 60, 1164–1174
13 Domini, F. and Braunstein, M.L. (1998) Recovery of 3D structure from motion is neither Euclidean nor affine. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1273–1295
14 Domini, F. and Caudek, C. (1999) Perceiving surface slant from deformation of optic flow. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 426–444
15 Domini, F. and Caudek, C. (2003) Recovering slant and angular velocity from a linear velocity field: Modeling and psychophysics. *Vision Res.* 43, 1753–1764
16 Gibson, J.J. (1979) *Ecological Approach to Visual Perception*, Erlbaum
17 Koenderink, J.J. (1986) Optic flow. *Vision Res.* 26, 161–180
18 Tittle, J.S. *et al.* (1995) The systematic distortion of perceived 3D structure from motion and binocular stereopsis. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 663–678
19 Liter, J.C. *et al.* (1993) Inferring structure from motion in two-view and multiview displays. *Perception* 22, 1441–1465
20 Hogervorst, M.A. and Eagle, R.A. (2000) The role of perspective effects and accelerations in perceived three-dimensional structure-from-motion. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 934–955
21 Domini, F. *et al.* (1997) Misperceptions of angular velocities influence the perception of rigidity in the kinetic depth effect. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 1111–1129

22 Vanduffel, W. *et al.* (2002) Extracting 3D from motion: differences in human and monkey intraparietal cortex. *Science* 298, 413–415
23 Orban, G.A. *et al.* (1999) Human cortical regions involved in extracting depth from motion. *Neuron* 24, 929–940
24 Morrone, C. *et al.* (2000) A cortical area that responds specifically to optic flow, revealed by fMRI. *Nat. Neurosci.* 3, 1322–1328
25 Sugihara, H. *et al.* (2002) Response of MSTd neurons to simulated 3D orientation of rotating planes. *J. Neurophysiol.* 87, 273–285
26 Lagae, L. *et al.* (1994) Responses of macaque STS neurons to optic flow components: a comparison of areas MT and MST. *J. Neurophysiol.* 71, 1597–1626
27 Cornilleau-Pérès, V. and Gielen, C.C.A.M. (1996) Interactions between self-motion and depth perception in the processing of optic flow. *Trends Neurosci.* 19, 196–202
28 Wexler, M. *et al.* (2001) Self-motion and the perception of stationary objects. *Nature* 409, 85–88
29 Peh, C.H. *et al.* (2002) Absolute distance perception during in-depth head movement: calibrating optic flow with extra-retinal information. *Vision Res.* 42, 1991–2003
30 Hildreth, E.C. *et al.* (1990) The perceptual buildup of three-dimensional structure from motion. *Percept. Psychophys.* 48, 19–36
31 Grzywacz, N.M. and Hildreth, E.C. (1987) Incremental rigidity scheme for recovering structure from motion: position-based versus velocity-based formulations. *J. Opt. Soc. Am. A* 4, 503–518
32 Caudek, C. and Domini, F. (1998) Perceived orientation of axis of rotation in structure-from-motion. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 609–621
33 Domini, F. *et al.* (1998) Discriminating constant from variable angular velocities in structure from motion. *Percept. Psychophys.* 60, 747–760
34 Caudek, C. and Rubin, N. (2001) Segmentation in structure from motion: modeling and psychophysics. *Vision Res.* 41, 2715–2732
35 Domini, F. *et al.* (2003) Temporal integration of motion and stereo cues to depth. *Percept. Psychophys.* 65, 48–57
36 Caudek, C. *et al.* (2002) Short-term temporal recruitment in structure from motion. *Vision Res.* 42, 1213–1223
37 Domini, F. *et al.* (2002) Temporal integration in structure from motion. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 816–838